

Technical Report 1121

Question Generation as a Learning Multiplier in Distributed Learning Environments

Arthur C. Graesser

University of Memphis

Consortium Research Fellows Program

Robert A. Wisher

U.S. Army Research Institute

October 2001



**United States Army Research Institute
for the Behavioral and Social Sciences**

Approved for public release; distribution is unlimited.

20020306 127

**U.S. Army Research Institute
for the Behavioral and Social Sciences**

A Directorate of the U.S. Total Army Personnel Command

**EDGAR M. JOHNSON
Director**

Technical Review by

James Dees, HQ U.S. Army Training and Doctrine Command
Joseph Psotka, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: TAPC-ARI-PO, 5001 Eisenhower Ave., Alexandria, VA 22333-5600.

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) October 2001		2. REPORT TYPE Interim		3. DATES COVERED (from... to) May 2000 to March 2001	
4. TITLE AND SUBTITLE Question Generation as a Learning Multiplier in Distributed Learning Environments				5a. CONTRACT OR GRANT NUMBER	
				5b. PROGRAM ELEMENT NUMBER 20363007	
				5c. PROJECT NUMBER A792	
6. AUTHOR(S) Arthur C. Graesser (University of Memphis) and Robert A. Wisher (U.S. Army Research Institute)				5d. TASK NUMBER 208	
				5e. WORK UNIT NUMBER	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-II 5001 Eisenhower Avenue Alexandria, VA 22333-5600				8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: TAPC-ARI-II 5001 Eisenhower Avenue Alexandria, VA 22333-5600				10. MONITOR ACRONYM	
				11. MONITOR REPORT NUMBER Technical Report 1121	
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.					
13. SUPPLEMENTARY NOTES					
14. ABSTRACT (Maximum 200 words) This report provides a rationale for question generation as a workable learning multiplier in distributed learning environments. The rationale was derived from a thorough review of recent research on questioning from multiple perspectives: psychology, cognitive science, computational linguistics, and information systems design. Based on this review, nine practices were identified for immediate use in both the conventional classroom and distributed learning settings. If employed properly, question generation strategies in distributed learning can increase a soldier's depth of understanding about the workings of a complex system. The strategy is particularly useful for asynchronous distance learning, where the instructor is not necessarily available to answer questions promptly.					
15. SUBJECT TERMS training, education, learning, distributed learning, distance learning, online instruction, questions, knowledge representation					
SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT Unclassified	20. NUMBER OF PAGES 42	21. RESPONSIBLE PERSON (Name and Telephone Number) Dr. Robert A. Wisher (703) 617-5540
16. REPORT Unclassified	18. ABSTRACT Unclassified	18. THIS PAGE Unclassified			

Technical Report 1121

Question Generation as a Learning Multiplier in Distributed Learning Environments

Arthur C. Graesser
University of Memphis
Consortium Research Fellows Program

Robert A. Wisher
U.S. Army Research Institute

Advanced Training Methods Research Unit
Franklin L. Moses, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

October 2001

Army Project Number
2O363007A792

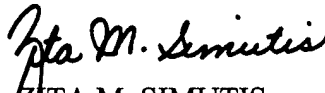
Personnel Performance
and Training

Approved for public release; distribution unlimited.

FOREWORD

The U.S. Army Research Institute is examining the use of distance learning technologies for use by soldiers in an “on demand” environment. The research under the WEBTRAIN project, sponsored by the Training and Doctrine Command (TRADOC), seeks to provide guidance to the U.S. Army as it transforms from a classroom-centric method of instruction to one that is more soldier-centric and asynchronous.

An important part of learning is asking questions. How this can be accomplished in learning environments that are distributed over time and distance is a vital concern to course designers and distance learning providers. As the Army shifts its focus to an anytime, anywhere training paradigm, the learner’s need to pose specific questions and receive accurate feedback will persist. The potential to automate such a function, in the absence of live instructors, needs to be examined. The methods and results of research from a range of disciplines, psychology, cognitive science, linguistics, and information system design, offer application areas for advanced research in Army training settings. These application areas were presented to the Training Development and Analysis Directorate, TRADOC, on 12 July 2001.


ZITA M. SIMUTIS
Technical Director

Acknowledgement

The authors would like to thank Dr. Michael Singer for comments on an earlier draft of this report. The assistance of Scotty Craig, Brent Olde, Victoria Pomeroy, and Sonja Rajan in identifying articles for review is sincerely appreciated. We would also like to thank Kara Orvis for her detailed review of the manuscript.

Question Generation as a Learning Multiplier in Distributed Learning Environments

EXECUTIVE SUMMARY

Research Requirement:

As the Army transforms its training enterprise to a model that relies more on learner-centric instruction, soldiers will assume increased responsibility for acquiring knowledge and developing skills. Questions by learners arise as a natural part of the learning process. Questions indicate an individual's lack of specific knowledge, an inability to comprehend, or a need to construct meaning. Responsive answers to well-structured questions can have a powerful effect on learning. Unfortunately, anytime-anywhere instruction, particularly asynchronous instruction, may lack the mechanism for students to get a timely response to a spontaneous question. An examination of the research literature on question generation and ideas for incorporating best practices into the future Army learning architecture are needed.

Procedure:

A review of the literature on question generation as an important learning mechanism from both a cognitive and information sciences perspective was conducted. Then, empirical studies of question generation in different learning environments using various information technologies were examined. How the benefits of question generation may be applied to distributed learning in military training environments was then assessed, with a specific objective to identify application areas for implementation.

Findings:

More than 100 documents were reviewed. There was ample evidence that training students to ask good questions can improve the comprehension and learning of technical materials. Based on the review, nine application areas for implementing the research into productive practices were identified.

Utilization of Findings:

This report is relevant to planners and training developers charting the future course of distributed learning and to software designers dealing with the challenge of incorporating question generation into advanced distributed learning systems. Two of the practice areas identified, presenting examples of good questions and having groups generate questions collaboratively, are already being pursued by the U.S. Army Research Institute with U.S. Army Training and Doctrine Command as part of an effort called Army TEAMThink.

QUESTION GENERATION AS A LEARNING MULTIPLIER IN DISTRIBUTED LEARNING ENVIRONMENTS

Contents	Page
Introduction.....	1
Question Generation	2
The Impact of Question Generation on Learning	2
Question Generation Mechanisms	5
Knowledge Representations and Cognitive Processes	8
Question Taxonomies	10
Multiple Choice Question Formats.....	12
Empirical Studies of Question Generation	14
Depth of Questions	14
Collaborative Question Generation	15
Practices for Increasing the Frequency and Quality of Questions	16
Practice 1: Clarify the learning objectives and test criteria	16
Practice 2: Present challenges that create cognitive disequilibrium	17
Practice 3: Give feedback on particular comprehension deficits.....	17
Practice 4: Present examples of good questions	17
Practice 5: Present generic question frames	17
Practice 6: Use conversational agents to scaffold the construction of questions.....	17
Practice 7: Have groups generate questions collaboratively	18
Practice 8: Concretize the author	19
Practice 9: Use computer-generated critiques of questions	19
Application Areas	19
Summary	23
References.....	25
APPENDIX A Question Taxonomy Proposed by Graesser and Person (1994).....	A-1
APPENDIX B TRADOC Guidelines for Preparing Multiple Choice Test Item.....	B-1

LIST OF TABLES

Table 1 Possible reasons why question generation learning (QGL) might improve learning.	3
Table 2 Assumptions behind sincere information-seeking questions (Van der Meij, 1987).	6
Table 3 Important types of knowledge representation for the military.	10
Table 4 Types of cognitive processes that are relevant to questions.	10

Question Generation as a Learning Multiplier in Distributed Learning Environments

Introduction

In tomorrow's dynamic threat environment, Army forces may have to deploy on short notice, to initiate operations that cannot be adequately predicted, and to conduct operations that have not been practiced. Future staffs must be able to interact with globally distributed expertise. Future forces must be highly adaptive and be able to reorganize quickly to meet threats effectively. Soldiers must continually learn, and rehearse, whether in school, at home station, at home, or in the theater of operations. Providing instruction on an anytime-anywhere basis is a key to improving performance and maintaining military readiness.

The Army Distance Learning Program (TADLP), initiated in 1996, has the vision to deliver individual, collective, and self-development training to soldiers and units anytime and anywhere through the application of multiple means and technologies (TRADOC, 1999). More than 500 courses are programmed for distance learning redesign. Through this transformed content, soldiers can utilize self-paced distance learning modules delivered at their home station, at their workplace, or in their own residence. Soldiers will have an increased responsibility for acquiring and maintaining skills, often in the absence of a face-to-face instructor. Within this paradigm shift, the role of the instructor is likely to change as the learning process becomes more soldier-centric.

Whether instruction is delivered online or in the classroom, questions arise as a natural part of the learning process. Students may request an explanation, a clarification, a concrete example, or make some other form of inquiry that reflects uncertainty. Questions are usually learner-centric, indicating an individual's lack of specific knowledge, an inability to comprehend, or a need to construct meaning. Unfortunately, anytime-anywhere instruction may lack the mechanism for students to get a timely response to a spontaneous question. This is particularly true of around-the-clock asynchronous learning environments in which omnipresent instructors are impractical due to the expense it would entail.

A fundamental issue, then, concerns the impact of question asking on learning. How does the process of question asking influence learning, performance, and student satisfaction? There is ample research evidence that training students to ask good questions can improve the comprehension and learning of technical materials (Palincsar & Brown, 1984; King, 1992). How can this finding be beneficially applied to advanced distributed learning environments? A second issue concerns the nature and frequency of the questions that are likely to be asked. To what extent can these questions be anticipated beforehand? What automated mechanisms can be developed to answer them? How important is the capability to respond quickly rather than after an unpredictable waiting period?

This report focuses on the application of question generation to soldier-centric distributed learning (DL) environments. It is intended for a broad audience. For planners and training developers charting the future course of distributed learning, it proposes application areas appropriate for "building out" the research into productive practices. For researchers and software designers dealing with the challenge of incorporating question generation into advanced

training systems, it offers a rationale from both a cognitive and information sciences perspective. The report begins with a review of the cognitive underpinnings of question generation. This first section covers the cognitive mechanisms of question generation as an important learning mechanism. The primary focus is on sincere information-seeking (SIS) questions, as opposed to questions that merely monitor the flow of conversation. The second section covers empirical studies of question generation in different learning environments using various information technologies. We present preliminary results of an experiment on soldiers' generating questions for other soldiers to answer, all through a Website. The third section discusses methods for increasing the frequency and quality of questions. The fourth section presents some schemes to identify and predict "frequently asked questions" (FAQ's) in asynchronous learning environments and some methods to automate the handling of these questions. In the fifth section, we recommend a number of application areas for test and implementation.

Question Generation

The design of question asking and answering facilities in any information system requires an in-depth analysis of question generation mechanisms (Lauer, Peacock, & Graesser, 1992). Such mechanisms specify the representation of the subject matter, the cognitive processes associated with inquiry, and the social context of the communicative interaction. There are many reasons why a question facility can fail in a learning environment or some other information system. This section should help researchers, software designers, and course developers identify some of the potential barriers while appreciating some of the salient advantages of sophisticated question generation facilities. There will be an emphasis on learning environments because the long-term objective of this report is to facilitate learning in Army training efforts, notably TADLP efforts. However, this section provides relevant insights to the design of any information system.

The Impact of Question Generation on Learning

Researchers in cognitive science and education have often advocated learning environments that encourage students to generate questions (Beck, McKeown, Hamilton, & Kucan, 1997; Dillon, 1988; Pressley & Forrest-Pressley, 1985). There are several reasons why question generation might play a central role in learning. The most frequent reason is that it promotes active learning and construction of knowledge (Bransford, Goldman, & Vye, 1991; Brown, 1988; Scardamalia & Bereiter, 1985; Papert, 1980). The learner needs to actively construct knowledge during learning rather than being bombarded with a large volume of information. Therefore, a computerized learning environment should be a scaffold for active construction of knowledge (including answering student questions) rather than being a mere information delivery system (Edelson, Gordin, & Pea, 1999; Papert, 1980; Schank, 1999). Constructivist approaches have been so compelling during the last decade that they have shaped the standards for curriculum and instruction in the United States during the last decade, e.g., *Standards for the English Language Arts* (NCTE, 1996), *Curriculum and Evaluation Standards for School Mathematics* (NCTM, 1989), *National Science Education Standards* (NRC, 1996). Most of these standards apply to education (i.e., reading, mathematics, science) rather than the training of specific skills and content, but there is no principled reason for doubting the utility of constructivism in training.

Question-generation learning. Question generation-learning is an environment in which learners are encouraged or compelled to ask questions while they study material. That is, they ask questions after trying to comprehend each sentence, paragraph, or section in the text. Answers to the questions may be provided by an expert immediately (as in synchronous distance learning) or after a delay (as in asynchronous distance learning). Alternatively, there might be no satisfactory answers to the questions. Question-generation learning (QGL) may be effective for several reasons, over and above the answers that are delivered in an ideal learning system. The potential advantages of QGL are summarized in Table 1. Available research has not dissected which of the particular components in Table 1 are primarily responsible for the potential benefits of QGL.

Table 1

Possible reasons why question-generation learning may improve learning.

-
- (1) **Active learning.** Learners actively construct knowledge in the service of questions rather than passively receiving information.
 - (2) **Metacognition.** Learners become sensitive to their own knowledge deficits and comprehension failures while they attempt to comprehend the material.
 - (3) **Self-regulated learning.** Learners take charge of both identifying and correcting comprehension problems.
 - (4) **Motivation and engagement.** Learners are more motivated and engaged in the material because the learning experience is tailored to their own needs.
 - (5) **Building common ground with author.** Learners achieve more shared knowledge with the author of the material.
 - (6) **Transfer appropriate processing.** Learners are normally tested by answering questions, so generating questions as part of the learning process should improve the overlap between comprehension representations and test representations.
 - (7) **Coding of cognitive representation.** The cognitive representations are more precise, specific, and elaborate when learners generate questions.

It is well documented that improvements in the comprehension, learning, and memory of technical material can be achieved by training students to ask questions during learning (Ciardiello, 1998; King, 1989, 1992, 1994; Rosenshine, Meister, & Chapman, 1996). The process of question generation accounts for a significant amount of these improvements from QGL, over and above the information supplied by answers. Moreover, QGL is most effective when students are trained to ask *good* questions. Exactly what constitutes a good question will be discussed later when a taxonomy of questions is presented. Rosenshine et al. (1996) provided the most comprehensive analysis of the impact of QGL on learning in their meta-analysis of 26 empirical studies that compared QGL to conditions with appropriate controls. Their meta-analysis included 17 studies in which students were taught question-asking skills and later tested on their comprehension or memory. The other nine studies adopted a reciprocal teaching method in which the teacher spends 75% of the time asking and responding to questions; that is, the teacher models good question asking and then has the students ask the questions in a modeling-scaffolding-fading process. The outcome measures in these studies included standardized tests, short-answer or multiple-choice questions prepared by experimenters, and summaries of the texts. The median effect size was .36 for the standardized tests, .87 for the experimenter-

generated tests, and .85 for the summary tests. There were no significant differences between the 17 studies that had question asking instructions and the nine that had the reciprocal teaching method.

One informative result of the Rosenshine et al. meta-analysis was that the question format was important when training the learners how to ask questions. The analysis compared training with signal words (*who, what, when, where, why, and how*), training with generic question stems (*How is X like Y?, Why is X important?, What conclusions can you draw about X?*), and training with main idea prompts (*What is the main idea of paragraph X*). The generic-question stems were the best perhaps because they give the learner more direction, are more concrete, and are easier to teach and apply. This result is informative because it will give us some guidance in designing question-prompt capabilities for distributed learning applications.

Another informative result pertains to the feedback, comments, and corrections that learners receive on their questions. Most of the studies included in the meta-analysis probably provided feedback to the learners on the quality of their questions, but a description of such feedback was absent in most of the reported studies. Therefore, the role of question-asking feedback is an issue for future investigations, perhaps in basic research. Feedback can potentially come from a teacher, a peer learner, or a computer. MacGregor (1988), for example, developed a computer-based instructional format that modeled good questions when the learner appeared to be facing problems in question asking.

Question Generation in Distance Learning. The available research on QGL has not evolved to the point of having systematic comparisons between synchronous DL, asynchronous DL, and other learning environments. This comparison is presumably important because learners receive an immediate answer to their questions in synchronous learning but must wait several hours or days for an answer in asynchronous learning. We performed a literature search for relevant articles on QGL that were published during the last five years in the following sources: *ERIC, PsychLit, American Educational Research Journal, American Journal of Distance Education, Cognition & Instruction, Educational Researcher, Educational Technology Review Interactive Learning Environments, Journal of Educational Multimedia and Hypermedia, Journal of Educational Psychology, Journal of Experimental Education, Journal of Interactive Learning Research, Journal of the Learning Sciences, Review of Educational Research, and WebNet Journal*. (We hereafter refer to this as the "QG literature search"). We could not find a single study that compared questions asked in synchronous versus asynchronous DL or between DL and other learning environments.

More global social networks of users have been participating in computer supported collaborative learning (Otaga, Sueda, Furugori, & Yani, 1999; Songer, 1996) and informal peer-help networks (Greer, 1999). Users are distributed at different geographical locations and communicate by e-mail. Learners ask and answer questions of their peers, or communicate with designated experts. In some systems, participants in the network score points, or accumulate credits, by helping others and answering questions posed by their peers. The selfish participants who seek help, but never provide help, end up depleting their credits and eventually lose access to the community resource. One practical research issue to explore is whether learners will seek help more than supply help in the information economy. Also, the pragmatics of e-mail

communication may be quite different from face-to-face (FTF) communication. In FTF communication, questioners can routinely rely on an answer being supplied by the listener. Indeed, it would be rude not to respond to the questions. Another practical research issue is whether peers or experts are consulted most often when a learner has a question. A comfortable chat may be higher in priority than gaining quality information.

The conclusion that QGL is effective is indisputable, but there are two general concerns that might limit the scope of QGL in promoting learning gains. The first concern is that it is difficult to disentangle (a) the process of asking questions, (b) feedback to the learner on the quality of the question, and (c) the quality of the answer. The comparative impact of these three components on learning gains has not been reported in the available meta-analyses. The second concern is that the pragmatic context of question asking has been unnatural or vague in many of these studies. For example, consider the studies in which the learner is instructed to generate questions while reading. There is no obvious recipient of the question, so it is difficult to determine whether a question is a sincere information-seeking (SIS) question or is a forced exercise in elaborating the material. Consider the studies in which the teacher models question-asking skills. Most of these questions are not SIS questions because the teacher already knows the answers to the questions. What is needed is an authentic context in which learners are vested in the questions that they ask. The next subsection on question-asking mechanisms will address the pragmatics of question asking and the conditions in which *bona fide* questions are asked.

Question Generation Mechanisms

It could be argued that any given task that a person performs can be decomposed into a set of questions that a person asks and answers. For example, when a soldier encounters a device that malfunctions, the relevant questions are "What's wrong?" and "How can it be fixed?" When an officer reads a situation report, the relevant questions are "Why is this important?" and "What should I do about it, if anything?". When a young adult reads recruiting material, the relevant questions are "What's interesting?", "Do I want to join?", and "What are the perks?". The cognitive mechanisms that trigger question-asking and exploration patterns need to be understood in order to optimize learning in the modern workplace, whether it be text, visual displays, mechanical devices, electronic equipment, or telecommunication systems.

There are differences between SIS questions and questions that do not invite answers that the questioner particularly cares about (Graesser & Person, 1994; Kreuz & Graesser, 1993; Van der Meij, 1987). Van der Meij identified 11 assumptions that need to be in place in order for a question to count as a SIS question. These 11 assumptions are listed in Table 2. A question is a misfire (non-SIS question) if one or more of these assumptions are not met. For example, when a computer science teacher grills a student with a question in a classroom (*What is RAM?*), it is not a SIS question because it violates assumptions 1, 5, 8, and 10. When a lawyer cross-examines a witness with a question, it is not a SIS question because it violates assumptions 1, 3, 4, 5, and 8. A standard lawyer maxim is "never ask a question unless you know the answer." Similarly, assumptions in Table 2 get violated when there are rhetorical questions (*When does a person know when he or she is happy?*), gripes (*When is it going to stop snowing?*), greetings (*How are you?*), and attempts to redirect the flow of conversation in a group (a hostess asks Bill, who has been quiet all night, *So when is your next vacation Bill?*). In contrast, a question is a

SIS question when a person's computer is malfunctioning and the person asks a technical assistant: *What's wrong with my computer?*

Table 2

Assumptions behind sincere information-seeking questions (Van der Meij, 1987).

-
1. The questioner does not know the information he asks for with the question.
 2. The question specifies the information sought after.
 3. The questioner believes that the presuppositions to the question are true.
 4. The questioner believes that an answer exists.
 5. The questioner wants to know the answer.
 6. The questioner can assess whether a reply constitutes an answer.
 7. The questioner poses the question only if the benefits exceed the costs.
 8. The questioner believes that the respondent knows the answer.
 9. The questioner believes that the respondent will not give the answer in absence of a question.
 10. The questioner believes that the respondent will supply the answer.
 11. A question solicits a reply.

These pragmatic assumptions have nontrivial implications from the standpoint of the design of future advanced distributed learning systems. There are many ways that a learning system or some other type of information system can break down. Learners will quickly give up using a system if the learner is unable to articulate the information in sufficient detail to get useful answers (assumption 2), if the system does not correct misleading presuppositions (assumption 3), if the system does not supply much useful information in the answer (assumption 8), if the system has trouble delivering any information (assumptions 10 and 11), and if the system cannot recognize silly questions posed by the user (assumptions 1, 4, 5 and 7).

A cognitive computational model of question-asking, called PREG, has recently been developed by Graesser and his colleagues (Graesser, Olde, Pomeroy, Whitten, Lu, & Craig, in press; Otero & Graesser, in press). According to the PREG model, cognitive disequilibrium drives the asking of SIS questions (Collins, 1988; Festinger, 1957; Schank, 1999). That is, questions are asked when individuals are confronted with obstacles to goals, anomalous events, contradictions, discrepancies, salient contrasts, obvious gaps in knowledge, expectation violations, and decisions that require discrimination among equally attractive alternatives. The answers to such questions are expected to restore equilibrium (homeostatic balance). Otero and Graesser (in press) developed a set of production rules that specify the categories of questions that are asked under particular conditions (i.e., content features of text and knowledge states of individuals). Some example production rules are shown below.

(A) Unknown word. A reader may be ignorant of the meaning of a word.

IF A content word W (noun, main verb, or adjective) in the text is not known

THEN Ask: "What does W mean?"

(B) Unknown referent. The explicit text mentions a noun or pronoun N, but it is difficult to construct or identify a referent in the mental model that corresponds to N.

IF A referent of a noun or pronoun N is not known

THEN Ask: "What/which N?"

(C) Discrepant Statement. An explicit statement S in the text is discrepant with a reader's knowledge of the prior explicit text or with world knowledge.

IF Statement S clashes with prior text or with world knowledge & no relation in the explicit text is linked to and accounts for S

THEN Ask: "Why did S occur/exist?", "How did S occur/exist?", or
"Why does the author say S?"

There are dozens of these production rules, but it is beyond the scope of this report to present the full inventory of the knowledge structures and production rules. From the present standpoint, these rules provide an *a priori* theoretical foundation for generating frequently asked questions (FAQ) and deciding what queries to include in system design. More will be said about FAQ in a later section.

The ability to detect cognitive disequilibrium at appropriate points while reading is an excellent index of whether a person understands technical text (Baker, 1985; Burbules & Linn, 1988; Glenberg, Wilkinson, & Epstein, 1982; Kintsch, 1998; Otero & Campanario, 1990). For example, if there is a direct contradiction in the text, this should be noticed. Poor comprehenders gloss over such contradictions and sustain an "illusion of comprehension." A deep comprehender actively seeks possible contradictions, clashes with world knowledge, and gaps in background knowledge (Beck et al., 1997; Hacker, Dunlosky, & Graesser, 1998). The detection of cognitive disequilibrium can be manifested in several ways. Reading time slows down. Eye movements regress to previous sections of text, or between contradictory constituents. And of course, one obvious manifestation is that the learner asks questions (Graesser & McMahan, 1993; Otero & Graesser, in press).

Detecting Disequilibrium. The detection of cognitive disequilibrium is alone not sufficient for question generation. According to the research conducted by Graesser and McMahan (1993), the potential question asker must pass two additional hurdles after the detection of disequilibrium: articulation of the question in words (called verbal coding) and the courage to express the question in a social setting (called social editing). Thus, three stages need to be intact for a question to be produced (disequilibrium detection, verbal coding, and social editing). Graesser and McMahan investigated this by having college students read different versions of stories and mathematical word problems: contradictions between text statements, deletion of critical information, insertion of irrelevant information, and control (no anomalies). The likelihood of generating questions was higher in the anomaly conditions than in the control condition when subjects were instructed to generate questions, as would be predicted by the PREG model. It is informative to note that the likelihood that students asked questions was extremely low (.04) when they were not instructed to ask questions, but were merely permitted to ask questions of an experimenter in an adjacent room. Thus, questions do not surface when it is

a physical or social effort to ask them. It is important to minimize these barriers in learning and information systems.

In a project funded by the Office of Naval Research, Graesser and his colleagues investigated the questions that college students ask when an everyday device malfunctions (Graesser, Olde, Pomeroy, Whitten, Lu, & Craig, in press; Graesser, Olde, & Lu, 2001). After reading about a device (e.g., cylinder lock, dishwasher, electric bell, toaster), the participants received scenarios in which the device breaks down (e.g., *the key turns but the bolt doesn't move*, in the context of a cylinder lock) and they generated questions about the malfunction. The breakdown scenario was expected to put the students in cognitive disequilibrium. Deep comprehenders are expected to explore the faults of the breakdown, to ask good questions, and thereby to restore cognitive equilibrium. Eye-tracking data were also collected during question asking in one of the empirical studies. There are two straightforward predictions of the PREG model. First, those participants who have a deep understanding of the device should ask good questions that converge on faults. Second, the eye movements of deep comprehenders should converge on likely faults that explain the breakdown. Both of these predictions were supported in the empirical investigations. It is interesting to note that deep comprehenders do not necessarily ask more questions. They ask better questions but not a higher frequency of questions (Fishbein, Eckart, Lauer, van Leeuwen, & Langmeyer, 1990; Graesser et al., in press).

One practical implication of the PREG model is that it is necessary to put the learners in cognitive disequilibrium before they start to ask questions in which they have a vested interest. Otherwise, they settle for a shallow understanding that maintains an illusion of comprehension. Therefore, the learning environment should present challenges, obstacles, contradictions, puzzles, tough decisions, and other events that crack the barrier of shallow knowledge. One of the sobering facts about education, training, and comprehension is that most students are not sensitive to their own comprehension deficits. Research on comprehension calibration has revealed that there is a low correlation (.1 to .4) between a learner's perception of how well they comprehend technical text, as measured by ratings, and how well they actually comprehend the text, as measured by objective tests (Glenberg & Epstein, 1985; Maki, 1998). Students learn the limits of their knowledge after they work on challenging problems and apply didactic knowledge in practical arenas. These are precisely the learning contexts that create cognitive disequilibrium and stimulate questions.

Knowledge Representations and Cognitive Processes

The contrast between shallow and deep knowledge is fundamental (Bloom, 1956; Bransford et al., 1991), so we will define it in this subsection. Shallow knowledge consists of explicitly mentioned ideas in a text that refer to: lists of concepts, a handful of simple facts or properties of each concept, simple definitions of key terms, and major steps in a procedure (not the detailed steps). Deep knowledge consists of coherent explanations of the material that fortify the learner for generating inferences, solving problems, making decisions, integrating ideas, synthesizing new ideas, decomposing ideas into subparts, forecasting future occurrences in a system, and applying knowledge to practical situations. Deep knowledge is presumably needed to articulate and manipulate symbols, formal expressions, and quantities, although some individuals can master these skills after extensive practice without deep mastery. Deep knowledge is essential for handling challenges and obstacles because there is a need to

understand how mechanisms work and to generate and implement novel plans. Explanations are central to deep knowledge, whether the explanations consist of logical justifications, causal networks, or goal-plan-action hierarchies. It is well documented that the construction of coherent explanations is a robust predictor of an adult's ability to learn technical material from written texts (Chi, deLeeuw, Chiu, & LaVancher, 1994; VanLehn, Jones, & Chi, 1992; Webb, Troper, & Fall, 1995).

Researchers in cognitive science, artificial intelligence, and discourse processes have specified different types and levels of knowledge representation in rich detail (Graesser, Gordon, & Brainerd, 1992; Kintsch, 1998; Lehmann, 1992). This subsection succinctly points out those distinctions that have a salient bearing on question generation in DL and that clarify the distinction between deep and shallow learning.

Text and Pictures. The representations of texts and pictures can be segregated into the levels of surface code, explicit propositions, mental models, and pragmatic interaction. The most shallow level is the surface code, which preserves the exact wording and syntax of the explicit verbal material. When considering the visual modality, it preserves the low-level lines, angles, sizes, shapes, and textures of the pictures. The explicit proposition representation captures the meaning of the explicit text and the pictures. A proposition contains a predicate (main verb, adjective, connective) that interrelates one or more arguments (noun-referents, embedded propositions). Examples of propositions are *the soldier repaired the computer* [repair(soldier,computer)], and, in the acronymic parlance of the Army Battle Command System¹, the AMDWS operator can pass messages to other ATCCS within the TOC [can pass (AMDWS operator (within TOC(other ATCCS)))]. At the deepest level, there is the *mental model* of what the text is about. For everyday devices, this would include: the components of the electronic or mechanical system, the spatial arrangement of components, the causal chain of events when the system operates, the mechanisms that explain each causal step, the functions of the device and device components, and the plans of agents who manipulate the system for various purposes. Finally, there is the *pragmatic communication* level that specifies the main messages that the author is trying to convey to the reader (or the narrator to the audience). Examples of purposes of reading are to explain how to repair equipment, to advertise a product, or to protect someone from a hazardous condition.

The *types* of representations are theoretically different from the *levels*. From the present standpoint, there are several types of knowledge representation affiliated with the explicit propositions and mental models. Table 3 lists some important types of knowledge representations that are important for military contexts. Specific question categories are associated with particular types of knowledge representation. Each of these types of knowledge become progressively deeper to the extent that they are more fine-grained (i.e., the grain size has high resolution) and have more complex interconnections among subcomponents (i.e., there are more relational links and more links that deviate from a strict hierarchy).

¹ AMDWS refers to Air Missile Defense Workstation; ATCCS refers to Army Tactical Command and Control System; TOC refers to Tactical Operations Center.

Table 3

Important types of knowledge representation for the military.

-
- Agents.** Organized sets of troops, units, organizations, countries, complex software units, etc.
Examples are organizational charts, friend-foe networks, and client-server networks.
- Class inclusion.** One concept is a subtype or subclass of another concept.
For example, an *M1A2* is-a *tank* is-a *weapon system*.
- Spatial layout.** Spatial relations among regions and entities in regions.
For example, *Eagle Base* is-in *Bosnia* is-in *Eastern Europe*. *Bosnia* is-south-of *Germany*.
- Compositional structures.** Components have subparts and subcomponents.
For example, *a computer has-as-parts a monitor, keyboard, CPU, and memory*.
- Procedures and plans.** A sequence of steps/actions in a procedure accomplishes a goal.
Examples are performance steps in disassembling and removing a mine, and the sequence of actions to create a command post filter in the Common Tactical Picture application.
- Causal chains and networks.** An event is caused by a sequence of events and enabling states.
Examples are the stages in firing an artillery round and the chain of events in a battle.
- Others.** Property descriptions, quantitative specifications, rules, mental states of agents.

Cognitive processes also vary in difficulty. Table 4 lists the major types of cognitive processes that are relevant to an analysis of questions (Standards of the Army Training Support Center, 2000; Bloom, 1956). According to Bloom's taxonomy of cognitive objectives, the cognitive processes with higher numbers are more difficult and require greater depth. Recognition and recall are the easiest, comprehension is intermediate, and classes 4-7 are the most difficult. It is debatable whether there are differences in difficulty among categories 4-7, so they are often collapsed into one category in most applications of Bloom's taxonomy.

Table 4

Types of cognitive processes that are relevant to questions.

-
- (1) **Recognition.** The process of verbatim identification of specific content (e.g., terms, facts, rules, methods, principles, procedures, objects) that was explicitly mentioned in the text.
 - (2) **Recall.** The process of actively retrieving from memory and producing content that was explicitly mentioned in the text.
 - (3) **Comprehension.** Demonstrating understanding of the text at the mental model level by generating inferences, interpreting, paraphrasing, translating, explaining, or summarizing.
 - (4) **Application.** The process of applying knowledge extracted from text to a problem, situation, or case (fictitious or real-world) that was not explicitly mentioned in the text.
 - (5) **Analysis.** The process of decomposing elements and linking relationships between elements.
 - (6) **Synthesis.** The processing of assembling new patterns and structures, such as constructing a novel solution to a problem or composing a novel message to an audience.
 - (7) **Evaluation.** The process of judging the value or effectiveness of a process, procedure, or entity, according to some criteria and standards.

Question Taxonomies

Schemes for analyzing the qualitative characteristics of questions have been proposed by researchers in education (Beck et al., 1997; Ciardiello, 1998; Dillon, 1988; Flammer, 1981),

psychology (Dillon, Golding, & Graesser, 1988; Graesser & Person, 1994), computational linguistics and artificial intelligence (Allen, 1983, 1995; Jurafsky & Martin, 2000; Lehnert, 1997; Schank, 1986; Webber, 1988). Rather than reviewing all of these schemes, we will adopt the taxonomy proposed by Graesser and Person (1994). The Graesser-Person taxonomy is both grounded theoretically in cognitive science and has been successfully applied to a large number of question corpora (such as human and computer tutoring, questions asked while using a new computer system, questions asked while comprehending text, questions raised in television news, and questions by letters to an editor). The taxonomy is presented in Appendix A. Later in this report we will apply the Graesser-Person taxonomy to a corpus of questions asked by Army officers in a Captains Career Course.

The 18 categories are defined according to the content of the information sought rather than on question signal words (*who, what, why, how, etc.*). The question categories can be recognized by particular generic question frames (which are comparatively distinctive), but not simply by signal words (which can be ambiguous). As discussed earlier, generic question frames are easier to teach to learners and produce more learning gains than do signal words (Rosenshine et al., 1996; King & Rosenshine, 1993). Appendix A includes a question category label and the generic question frame for each category. Concrete examples of these 18 categories can be found in previous publications (Graesser, Lang, & Horgan, 1988; Graesser, Person, & Huber, 1992).

Graesser and Person (1994) reported that some of the categories of SIS questions are associated with deep comprehension and cognitive processes. Specifically, they regarded question categories 10 through 15 as deep-reasoning questions because they tapped causal networks, planning, and logical justifications (i.e., answers to *why, how, what-if, and what-if-not*). They tested this by analyzing students' SIS questions in tutoring sessions on research methods and basic mathematics. SIS questions constituted 28% of the student questions. Trained judges classified student questions into the 18 categories with a high degree of reliability (Chronbach's alpha of .81 or higher). The proportion of SIS questions that were in question categories 10-15 was recorded. A separate set of trained judges classified questions on depth using five levels that map onto Bloom's taxonomy; they could do this with high reliability (Chronbach's alpha = .85). A depth score was determined for each student, which consisted of the proportion of SIS questions that were in the deeper levels of Bloom's taxonomy. There was a .64 correlation between depth score and the proportion of student questions that were deep-reasoning questions. Most of the student questions were shallow (70%) rather than deep (30%). The most frequent deep-reasoning questions were in the instrumental/procedural category (66% of the deep-reasoning questions). One important result of the study was that, overall, students rarely ask deep-reasoning questions (.30 deep questions X .28 SIS questions = .08 deep-reasoning questions, or 8%). This finding is compatible with one major conclusion in the present report: Getting learners to ask good questions is an uphill challenge.

The majority of student questions (72%) are not SIS questions. Instead, they are questions that simply verify what they already know (*Aren't I supposed to add these columns?, Doesn't a factorial design have two independent variables?*) or that address the social interaction between student and tutor (*Is it my turn?, What did you say?, When does this session end?*). We call these common ground questions and metacommunication questions, respectively. We expect that these common ground and metacommunication questions will emerge frequently in

synchronous DL, particularly when the learner cannot see the instructor. In contrast, we expect them to be less frequent in asynchronous DL systems, because there is a cost in having a delayed answer and because the benefits of such communications are not adequate to justify an unpredictable wait (see assumption 7 in Table 2). SIS questions should be more prevalent in asynchronous than synchronous DL since time to reflect and compose a question is not constrained.

One can imagine a 3-dimensional grid that captures the landscape of questions that could potentially be asked about any given subject matter. The three dimensions include the Type of Knowledge (see Table 3), the type of cognitive process (see Table 4), and the question category (see Table 5). Given the number of categories in each dimension, there are $7 \times 7 \times 18 = 882$ cells in the landscape of questions. Some of these cells are more prevalent than others in any given subject matter. For example, consider procedures, which are step-by-step recipes performed by people. The shallow questions include concept completion, definition, feature specification, and quantification. Deep questions include causal antecedents (Why did X occur?, How did X occur?, What caused X?), causal consequences (What-if?, What if-not? What are the consequences of X?), goal orientation questions (Why do you do X?), instrumental procedural (How do you do X?), and expectational (Why don't you do X?). One can ask what the next step is in the procedure (What happens after action X is performed?), what action to take if a condition exists (What do you do if X exists/occurs), what is the relative ordering of actions (X before/after/during Y vs. indeterminate), and what the functions of actions are (Why is it important to do step X?).

Learning objectives vary, so a particular course will skew the emphasis on some cells more than others. Shallow knowledge is sufficient in certain educational and military training contexts, although it should be recognized that shallow knowledge does not necessarily transfer well to real-world practical situations. For example, shallow learning is prevalent when many people learn about the components of a computer, the jargon, and the relevant acronyms. But this shallow knowledge will not help much when the person needs to apply the computer in a real world task or to repair the computer when it malfunctions. There are occasions when deep knowledge is necessary. Such knowledge involves a different distribution across the cells in the landscape of questions. In any event, we believe this landscape of questions will be useful to designers of tests and distributed learning environments because it will expand the horizons on what questions are possible. The vast majority of test questions are shallow in current educational and training practices (Martinez, 1999). Our hope is that the depth and scope of the questions will broaden with advanced research and application that ingrains question generation as a learning multiplier.

Multiple Choice Question Formats

Questions can appear in many different formats on tests. The most popular formats are multiple choice, true-false, matching, short answer completion, and essay. Multiple choice (MC) is the most frequent format and is routinely adopted in tests constructed by Educational Testing Service (such as the *Scholastic Aptitude Test*), the information technology training enterprise (such as the certification exams for Microsoft, Cisco, and Novell), and the military (such as the *Armed Services Vocational Aptitude Battery*). Essay questions are more useful than MC questions in diagnosing cognitive states and tracing cognitive processes (Martinez, 1999), but

properly constructed MC items can do an excellent job tapping complex thought and deep knowledge. There is a rich psychometric tradition in the construction of multiple choice items on such tests (*American Educational Research Association, American Psychological Association, & National Council on Measurement in Education*, 1985; Department of Defense, 1983; Downing & Haladyna, 1997; Messeck, 1987). This subsection discusses how to construct a good MC question because this can serve training and courseware developers as well as instructors.

For starters, it is important to distinguish between format and content. The format refers to the composition of the question elements, methods of circumventing comprehension difficulty, and methods to minimize “giving away the answer” to sophisticated guessers. The content refers to the level and type of knowledge tapped (see Table 3) and to cognitive processes (see Table 4). The following terminology is adopted when describing question format:

- (1) **Context.** Information that sets the stage and precedes the question stem.
- (2) **Stem.** The focal question, without the answer options.
- (3) **Key.** The correct answer among the set of alternatives.
- (4) **Distracters.** The incorrect options among the set of alternatives.

In most applications, there are four or five answer options and only one correct answer. It is recommended that the distracter items should vary in distractibility. One distracter should be the near miss. This is the most seductive distracter that reflects a common misconception that people have. The discrimination between the key and the near miss should reflect an important learning objective or pedagogical point rather than being arbitrarily subtle or merely tapping frivolous detail. The *thematic distracter* has content that is related to the topic at hand, but is not correct. A learner who quickly scanned the learning materials would have trouble discriminating the key, the near miss, and the thematic distracter. The *unrelated distracter* would seem reasonable to someone who never read the material, but is in fact unrelated to the lesson content.

Army Guidance. Mr. James Dees and Mr. Edward Tyler of the Army Training and Doctrine Command (TRADOC) provided guidelines to the Army Research Institute for constructing any test question and some guidelines for constructing MC test questions in particular. These guidelines are presented in Appendix B. It should be apparent that there are many constraints that need to be taken into consideration when writing a good MC question. The process is difficult and time-consuming when done correctly.

The difficulty of constructing MC questions perhaps explains one outcome of our literature search on QGL. We were interested in whether there was any empirical research that tested whether the process of learners generating MC questions might improve their comprehension of learning material. This approach is being pursued by the Army in conjunction with Athenium, LLC in the Army TEAMThink project, which will be discussed later. Our literature search came up empty, which suggests that a project that pursues this approach is innovative. It is also likely to produce learning gains, given the well-documented effects of QGL that were discussed earlier. However, one challenge in implementing such a project will be teaching learners how to ask MC questions. Scaffolding will be needed to guide the learner in

constructing questions that follow particular formats, constraints, and content features. Possible methods of scaffolding will be described later in this report.

Empirical Studies of Question Generation

There is an idealistic vision that learners are vigorous question generators who actively self-regulate their learning. They identify their own knowledge deficits, ask questions that focus on these deficits, and answer the questions by exploring reliable information sources. Unfortunately, this idealistic vision is an illusion. The sobering truth is that the vast majority of learners have trouble identifying their own knowledge deficits (Hacker et al., 1998) and ask very few questions (Dillon, 1988; Good, Slavings, Harel, & Emerson, 1987; Graesser & Person, 1994). Graesser and Person's (1994) estimate of available studies revealed that the typical student asks .1 question per hour in a classroom and that the poverty of classroom questions is a general phenomenon across cultures. The fact that it takes about nine hours for a typical student to ask one question in a classroom is perhaps not surprising because it would be impossible for a teacher to accommodate constantly 25-30 curious students. The rate of question asking is higher in other learning environments. An average student asks 26.5 questions per hour in one-on-one human tutoring sessions (Graesser & Person, 1994). Thus, when there is an attentive question answerer, the rate of student question asking goes up 250-fold.

The upper bound in the number of student questions per hour is 135, at least according to Graesser, Langston, and Bagget (1993). The latter estimate is based on college students interacting with *Point & Query* educational software. The only way that these students can learn about a topic (e.g., woodwind instruments) is by asking questions and by reading answers to their questions. To ask a question, the learner first points to a word, phrase, picture element, or other hotspot on the computer screen in a hypertext system. Second, a menu of relevant questions about the hotspot is presented to the learner (e.g., *What does X mean?*, *What does X look like?*). Third, the learner selects one of the question options. Fourth, the answer is presented and the learner reads the answer. This cycle continues until the learning session is finished. It is extremely easy for a learner to ask a question; it simply involves two clicks of a mouse, one to select the hotspot and the other to select the question. The *Point & Query* software is similar to some other menu-based question asking systems (Sebrechts & Swartz, 1991). In this ideal learning environment, the student asks over 1,000 times as many questions as in a classroom.

Depth of Questions

One other striking result is that most student questions are shallow, as we reported earlier. Student questions are normally shallow, short-answer questions that recycle through the content and interpretation of explicit material; they rarely tap the deeper levels of knowledge representation and cognitive processes. What the students do mirrors what the teachers do in classrooms. Only about 4% of teacher questions are deep questions (Kerry, 1987; Dillon, 1988). The computer has the potential to improve the quality in addition to the quantity of student questions. The quality of student questions might improve in *Point & Query* software that presents only good questions for the students to model. Graesser et al. (1993) found that 5 times as many deep-reasoning questions were asked when there were deep question options on the *Point & Query* question menu and the students were given a task that required deep reasoning. Craig, Gholson, Ventura, and Graesser (2001) presented college students conversational

computer agents that modeled good question asking behavior. That is, there were talking heads that asked questions with synthesized speech. A control condition had the talking heads deliver the same content with monologues. There were over twice as many deep-reasoning questions in the condition that had the agents model question asking than in the control condition. Recall for the content was also significantly higher in the modeling condition. These findings confirm the conclusions presented earlier that learning gains can be realized by training students how to ask questions.

Group question asking may provide an environment that enhances learning and question asking. For example, Aducci (1998) had groups of students generate questions that would appear on an examination. It was necessary to give some guidance to the students on how to construct test questions and how to conduct the group interactions. Each group contributed 1 or more questions on an examination. The participants were children and the topic was mathematics. It appeared that this method improved learning because the percentage of grades of C or better increased from 71% to 89% in a group of 30 students. Unfortunately, the evaluation was not systematic, so it is an open question whether group question asking is effective in promoting learning gains. It is difficult to determine whether any advantages could be attributed to question generation per se or to group learning. Researchers have periodically reported that group learning has several advantages over individual learning (Slavin, 1995; Johnson & Johnson, 1989; Kagan, 1994). Springer, Stanne and Donovan (1999) conducted a meta-analysis on 37 studies in postsecondary education, in the areas of science, mathematics, and engineering. The effect sizes for small group learning were .51 for learning achievement, .47 for persistence in the course, and .50 for attitude toward the learning experience, compared with individual learning. The impact on learning achievement was greater for instructor-made exams (.59) than standardized instruments (.33).

The low frequency and quality of learner questions have the potential to improve in synchronous DL. An instructor can immediately answer student questions in synchronous DL, whereas this is impractical in a classroom. However, empirical data on question asking in DL is conspicuously absent in the literature.

Collaborative Question Generation

The Army Research Institute (ARI) and TRADOC are developing a specialized version of TEAMThink, a novel question-based, collaborative learning application. TEAMThink, a licensed product of Athenium LLC, stimulates students to actively engage in the creation of questions through small group collaboration. The questions are posed to other students through a Website. The collaborative question asking is predicted to increase learning gains on the subject matter. In accordance with a Technology Transfer Memorandum of Agreement between ARI and TRADOC, the benefits of TEAMThink are being tested in three Army schools.

As a baseline measure, students in the Captains Career Course at the Engineer School, Fort Leonard Wood, Missouri, participated in writing questions in a four-item multiple choice format. Students composed one or two questions on either of two topics: the military decision making process and the scheme of engineer operations, each covered recently in the course. The questions were composed off line without collaboration. Of the 173 questions generated, only 2% were considered not doctrinally compliant and 5% were considered trivial. Through the

Athenium Website, a subset of 34 questions were posed to n=108 students (excluding those from Allied Nations). The overall performance was 68% correct. Categorizing the 173 questions to the Graesser and Person taxonomy (Appendix A), 45% were feature specification, 34% were instrumental/procedural, and 8% were concept completion. Eight other categories had less than 5% of the questions. Seven categories, primarily those tapping deeper knowledge, were never used. This baseline will later be compared to questions generated through Web-enabled, small group collaboration.

In summary, available empirical research supports a number of conclusions about question generation and learning. QGL is effective in promoting learning gains. However, learners need to be trained to ask question because they ask very few questions in most learning environments and most of the questions are low in quality (i.e., shallow rather than deep). There is no systematic research on collaborative question generation in groups and the resulting effect on learning. There is no systematic research on the impact of generating questions during comprehension on learning gains. MC questions have many constraints in format and content, so they are too difficult for students to generate without scaffolding of instruction. Research on question asking in DL is conspicuously absent.

Now that we have covered the theoretical and empirical research on question asking and learning, the focus will shift to the role of questions in future DL systems. The remaining sections should be regarded as “informed speculation.” It is informed because of the existing body of research reported in the first two sections. It is speculative, however, because research on questions is conspicuously absent in the current DL environments. Existing empirical data are so sparse and fragmented that we will have to rely on basic research and theory to support our speculations. The fact that systematic research on QGL in DL environments is so sparse is not surprising given that there have been few evaluations of DL environments with adequate controls and control groups (Wisher, Champagne, Pawluk, Eaton, Thornton, & Curnow, 1999).

Practices for Increasing the Frequency and Quality of Questions

Learners clearly need some scaffolding to assist them in asking questions. Progressively more scaffolding will be needed to encourage (a) any question at all, (b) deep questions, and (c) multiple choice questions. Deep knowledge and format constraints require additional help. In this section, we propose methods for increasing the frequency and quality of questions. The methods are in no particular order of importance.

Practice 1: Clarify the learning objectives and test criteria

The learner can evaluate what questions are relevant if they know these objectives and criteria. Without this clarity, there are barriers in the question generation stages of identifying knowledge deficits and social editing. Some learning objectives require only shallow and procedural knowledge, such as assembling a new computer and logging onto a network. The test is whether computers get assembled and the learner successfully logs onto the system. Other learning objectives require deep knowledge about causal mechanisms, as in the case of diagnosing and repairing equipment. Designers of learning environments should provide example test questions that appropriately map onto the relevant cells in the matrix of questions

appropriate to a training program.

Practice 2: Present challenges that create cognitive disequilibrium

SIS questions are triggered by cognitive disequilibrium, as discussed in an earlier section. The learner needs to be presented challenges in the form of obstacles to goals, anomalous events, contradictions, discrepancies, obvious gaps in knowledge, and decisions that require discrimination among equally attractive alternatives. Such challenges are most engaging when they are not too easy or too difficult, but are at the learner's zone of proximal development (Rogoff, 1990; Vygotsky, 1978). This can be accomplished by having the DL system track the learner's mastery of the material and present challenges that are tailored to the learner's profile (within the "zone"). The challenges consist of example problems to solve, procedural tasks to execute, and breakdown scenarios to fix.

Practice 3: Give feedback on particular comprehension deficits

Most learners do not have good comprehension calibration skills so they do not notice the knowledge deficits that would otherwise drive questions. Feedback on particular knowledge deficits ends up penetrating the "illusion of comprehension" and promoting deeper comprehension and also deeper questions. Intelligent tutoring systems (ITS) are capable of recognizing bugs, misconceptions, and knowledge deficits at a fine-grained level (Lesgold et al., 1992; Graesser, VanLehn, Rose, Jordan, & Harter, in press), so the ITS technology would be one source of implementing this method in TADLP.

Practice 4: Present examples of good questions

The learner can acquire good question asking skills by modeling good question askers. Example good questions, in whatever format, can be available in relevant cells in the landscape of questions. The learner can observe these question items by pointing to different hotspots on the graphical user interface (GUI) or in a help facility. Alternatively, pairs of avatars can exhibit good questions in virtual dialogues. These modeling approaches have proved effective in improving the quantity and quality of questions.

Practice 5: Present generic question frames

Generic question frames (e.g., *What does X mean?*, *How do you do X?*, see Table 5) guide the user in selecting and articulating questions. They are pitched at a somewhat more abstract level than actual questions, but there is a finite number of frames that are acquired by the learner. The Point & Query interface adopts these generic question frames and substantially improves the quantity and quality of learner questions. The learner discovers categories of good questions by examining the options on the question menu. The software designer can tailor the question options to fulfill the learning objectives and the subject matter constraints.

Practice 6: Use conversational agents to scaffold the construction of questions

Researchers have developed computer-generated, animated, talking heads that have facial features synchronized with synthesized speech and appropriate gestures (Cassell & Thorisson, 1999; Cohen & Massaro, 1994; Johnson, Rickel, & Lester, 2000). These conversational agents have been used as navigators on web pages, as narrators, as avatars, and as tutors in intelligent tutoring systems. For example, AutoTutor (Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & TRG, 1999) is a conversational agent that teaches students about introductory

computer literacy by holding a conversation. Dialog moves in AutoTutor include backchannel feedback (*Uh-huh, Okay*), pumps (*Tell me more*), prompts (*primary memory includes ROM and ____*), assertions, hints, corrections, summaries, questions, and a variety of other speech acts. A turn-by-turn dialog scaffolds the learner to articulate information. Steve (Rickel & Johnson, 1999) is an avatar in virtual reality that shows the learner how to operate equipment, answers learner questions, and offers suggestions.

Conversational agents could be developed to guide learners in articulating questions in various formats, including MC questions. For example, the agent could follow the following script that guides the user in generating a MC question:

- Step 1: Select a question from a cell in the landscape of questions.
- Step 2: Show an example question in a selected cell
- Step 3: Generate a question stem.
- Step 4: Generate the key.
- Step 5: Generate a near miss distracter.
- Step 6: Generate a thematic distracter.
- Step 7: Generate a remote distracter.

Each step would be augmented by a conversational finite state transition network (Graesser, Person, Harter, & TRG, 2000; Jurafsky & Martin, 2000) that allows learners to ask clarification questions (e.g., *What is a stem?*), that answers these questions, that makes suggestions (*At this point, you need to generate a stem*), that gives feedback (*uh-huh, that's right, okay*), and that gives hints (*Is there an important misconception that motivates this near miss?*). A conversational agent for question generation is well within the realm of current technologies.

Practice 7: Have groups generate questions collaboratively

In the TEAMThink project on group question generation, one student generates a MC question and key, whereas partners in a small group critique the question as well as the proposed distracters. Later on, the generated questions are used on tests with other small groups. Alternative role assignments are currently being explored in the project. For example, a third student might revise the question, in light of the feedback from the partner. So, there potentially could be a question asker, a critiquer, and a reviser. There also could be feedback from an expert composer before a final question is revised. One possible augmentation would be to have a competitive game in which groups compete in their generation of questions. Another performance-based assessment would be to have other learners, in other groups, answer the question and evaluate the question on discriminative validity (high performing students answering it correctly, low performing students missing it). Learners could receive feedback on the question quality according to psychometric indices.

There are other role assignments that might produce more questions, better questions, and/or more learning. One learner could generate the question stem, a second the key, a third the near miss, a fourth the thematic distracter, and a fifth the remote distracter. Each learner would also justify what is produced from the standpoint of content or question quality. This approach might sharpen important distinctions and yield deeper learning.

Practice 8: Concretize the author

The author of a document or lesson is normally invisible. Many readers assume that printed text is indisputable truth, rather than it being the best possible account that an author could prepare at a particular point in time. Beck et al. (1997) have designed a *Questioning the Author* method that trains students to imagine an author in flesh and blood and to question what the author says. Why does the author make a particular claims? What evidence is there? How does one idea in a text relate to another idea? Why would the author express something in a particular way? This Questioning the Author method improves questions, comprehension, and metacomprehension. So a picture of an author and example questions that challenge what the author says might be incorporated in DL applications.

Practice 9: Use computer-generated critiques of questions

Advances in computation linguistics, corpus linguistics, and natural language processing in artificial intelligence have made it feasible to design computer facilities that critique questions on quality. Some of these advances will be discussed in the next section. Questions can be automatically classified according to the Graesser and Person (1994) question categories (see Table 5). Once classified, they can be evaluated on depth in the critique.

Answer options in MC questions can be evaluated on correctness using latent semantic analysis (LSA) (Landauer, Foltz, & Laham, 1998; Foltz, Gilliam, & Kendall, 2000), a technique that will be discussed further in the next section. Landauer and Dumais (1997) created an LSA representation with 300 dimensions from 4.6 million words that appeared in 30,473 articles in *Grolier's Academic American Encyclopedia*. They submitted to the LSA representation the synonym portion of the TOEFL test, a test developed by the Educational Testing Service to assess how well non-native English speakers have mastered the words in the English language. The test has a four-alternative, forced choice format, so there is a 25% chance of answering the questions correctly. The LSA model selected the alternative that had the highest match with a comparison word. The LSA model answered 64.4% of the questions correctly, which is essentially equivalent to the 64.5% performance for college students from non-English speaking countries. There was also a significant correlation between the relative likelihood that the 3 distracters were selected and the likelihood that humans found them to be seductive lures. In another study, Foltz et al. (2000) created a 300 dimensional space that captured a text on introductory psychology and then submitted the LSA representation to the MC questions used in the class. The LSA space received a grade of C-. One could imagine having learners get feedback on their own MC questions by inspecting the likelihood that LSA selects each alternative. If the key had roughly the same LSA score as the near miss, then that would be a desirable feature, as long as there is a principled reason that the near miss is incorrect.

Application Areas

This section identifies several application areas for question generation as a learning multiplier in distributed learning environments. We will focus on distributed learning environments that are potentially useful to the military, including the reuse of sharable content through application of the Sharable Content Object Reference Model (SCORM).

Communication Channels and Media. One important consideration is the communication medium between questioner and answerer. In face-to-face (FTF) interaction, there is both temporal and spatial contiguity, so the speech participants co-experience the communication and the conversational setting. It is easy for the listener to give backchannel feedback (speech, head nods, gestures, facial expressions) while the speaker talks. Such feedback serves to repair the speaker's conversation, to fill in the speaker's words, to speak simultaneously, and to perform other collaborative acts that are well known in the speech and discourse literature (Clark, 1996; Fox, 1993). In FTF conversation, speakers routinely ask counter-clarification questions (*Which disk are you talking about?*) and produce metacognitive utterances (*I see. That makes sense. Do you follow? Okay?*). These questions solidify common ground between speech participants, a grounding that facilitated further by a shared physical environment. In FTF interaction, it is not normally necessary to produce metacommunication utterances (*Could you repeat that? Do you hear me? I'll say a few things now*) because there are virtually no barriers in the communication channel. In FTF interaction, there are comparatively few words per turn because it is easy to negotiate turn-taking through intonation, facial expressions, gestures, tight temporal coupling, and floor maintenance utterances (*umm, uhh, soo..*).

Alternative media present barriers that are absent in FTF interaction. Asynchronous DL is at the opposite end of the spectrum because the interaction is neither temporally nor spatially contiguous. Hillman (1999) has explored some of the differences in discourse patterns between these two media. In their analysis of computer classes, teachers spoke 73% of the sentences in FTF interaction but only 49% in asynchronous DL. This strongly suggests that students can be more active constructors of knowledge in asynchronous DL. Organizing information and the lesson is more explicit in DL than FTF because of the lack of nonverbal channels and the lack of immediacy of the interaction. However, Hillman did not analyze the discourse in the level of detail that is routinely conducted by researchers in the field of discourse processing. The discourse acts that should be infrequent in asynchronous DL are backchannel feedback, conversational repair, metacognitive utterances, and metacommunicative utterances.

Responses to such acts have comparatively low information value, whereas there is a high cost in the form of wait time and miscommunication (because of the absence of shared context); since the benefits do not exceed the costs, these questions would not be asked (see assumption 7 in Table 2). The number of words per turn should be much longer in asynchronous DL because of the need to create context and to optimize the information load per turn. Perhaps this problem could be minimized with static pictures of the instructor in addition to the materials, an expanded form of audiographics. This might work for person-to-person communication, but not when the learner is interacting with a generic trainer, and a trainer is interacting with a generic learner. It is interesting to point out that deep SIS questions are predicted to be more prevalent in asynchronous DL because there should be a high load of information per turn, there is more time to compose questions, and there is more time to reflect and plan during question composition. There should be fewer questions, but higher quality questions in asynchronous DL than FTF.

The media in between FTF and asynchronous DL present interesting research issues from the standpoint of question generation and discourse. Wisher et al. (1999) identified the various communication media and reported their usage and effectiveness in recent DL applications.

Clark's (1996) analysis of discourse and communication media provides an excellent theoretical guide for analyzing the different forms of synchronous DL. Speech participants need to monitor common ground while they communicate, so there needs to be feedback among them on what others think about what gets said. There are different degrees of feedback when speaker #1 expresses message M to speaker #2 (see also Graesser et al., 1999):

- (1) I am near you
- (2) I am listening
- (3) I hear you
- (4) I acknowledge that you are talking
- (5) I understand what you say
- (6) I understand what you say, and I'll demonstrate it by expanding upon it.
- (7) I agree with what you say
- (8) I have an attitude toward what you say

In FTF interaction, utterance number 1 can be accomplished by spatial proximity, number 2 by the listener facing the speaker, number 3 by facial expressions that respond to the other's speech, number 4 by any form of backchannel feedback, number 5 by head nodding or positive spoken utterances (*yeah, I see*), number 6 by informative assertions that are relevant to the topic, number 7 by explicit agreement or more dramatic forms than in number 5, and number 8 by emotional expressions, sarcasm, or humor.

These levels can be quickly accomplished in FTF interaction, but must be more overtly and explicitly expressed in other communication media (Vrasidas & McIsaac, 1999). For example, consider two-party telephone conversations. Number 1 is impossible but the others are routinely accomplished by the speech, pauses, and intonation patterns, but not by body positioning, facial expressions, gestures, and interacting with the setting. In conference calls with several people, there is uncertainty about who is present, so it is polite for everyone to introduce themselves. It is more difficult for everyone to reconstruct who the primary speakers are. There needs to be more metacommunicative utterances (*Can you hear me? Who is speaking? Can I say something?*) that manage the conversation and handle 2, 3, and 4. There needs to be more metacognitive utterances (*I'm not following?, Could you say that in English? Could I think about that for awhile?*). This is all accomplished with speech, pauses, and intonation contours. The features of the telephone artifact also play a role, of course. If there is a dial tone in the middle of the conference call, then that disables 2-8. A dial tone after a heated exchange satisfies 8. The conversational flow and understanding can be disrupted with poor audio quality, such as delays for transmission, delayed auditory feedback, and signal degradation that changes intonation parameters.

Adding the visual modality to the audio modality in synchronous DL is expected to make it easier for speech participants to communicate, even though such improvements do not necessarily improve learning (Wisher & Curnow, 1999). The review of DL by Wisher et al. (1999) reported that 57% of the DL research conducted in training environments was video teletraining with two-way video and two-way audio. The video channel opens up facial expressions, gestures, body positioning, and interactions with the setting. Consequently, we would expect fewer metacommunication and metacognitive questions when the video media are

added. One of the challenges has been to coordinate the timing between the speech and visual media. Many individuals get irritated when facial expressions, gestures, and mouth movements are out of synch with the speech. It is conceivable that deep SIS questions are more prevalent without the visual modality. Speakers do not have to sustain polite body postures and facial expressions when there is no visual media, so there are more cognitive resources available for deeper thinking. Research in neurolinguistics has revealed that when people are deep in some forms of thought, they try to remove visual stimuli by closing their eyes or looking up at the ceiling.

In the future, there needs to be more research that empirically investigates question generation while adults use these different media and communication channels. This section has identified some theoretical reasons for expecting differences, but the empirical research has yet to be conducted.

Reducing the barriers between questions and answers. An earlier subsection of this report discussed methods of minimizing barriers to question generation and maximizing the quantity and quality of questions. The present subsection addresses barriers between a question that gets asked and useful answers. A frequent complaint about contemporary information retrieval systems is that the performance of these systems is poor. It is difficult to pose questions that end up supplying useful information. Recently, at a text retrieval conference on question answering, sixteen systems were entered in a competition. A set of questions were asked, the query-retrieval systems attempted to fetch the answers from the web, and then the top five answers were examined for each questions. A recall score was defined as the likelihood that the correct answer was among the top five answers. Recall scores were moderately impressive (around 50%) for shallow questions with short answers (concept completion, quantification), but notoriously poor for deep questions that required long answers (*why, how*).

Quite clearly, learners will give up asking questions if the system fails to deliver useful information. If assumptions number 8, 10, or 11 in Table 2 are violated, then the learner will quickly give up using the tool. Instead, they will ask a human (who may deliver faulty information) or will give up asking any question at all. This was a painful lesson in our early assessments of AutoTutor (Graesser et al., 1999; Graesser, VanLehn, et al., in press), the tutoring system that we developed to teach college students computer literacy by holding a conversation with them. AutoTutor could handle definitional questions and verification questions, but not any of the other categories in Table 5. When AutoTutor couldn't handle a question, it answered *That's a good question, but I can't answer it*. This had to happen only once or twice before a student stopped asking any SIS questions altogether. Similarly, in the classroom, student questions may not be understood by the teacher, so the utility of the answers may be marginal; this may partially explain why question asking in some children quickly extinguishes shortly after they enter school. Users quickly evaluate whether an information system is delivering useful information, whether the information system is a web facility, a textbook, or a human expert. There needs to be an extremely high benefit-to-cost ratio for adults to stay motivated asking questions. One direction for the future is to substantially increase this ratio.

Reuse of questions. One goal of the Advanced Distributed Learning (ADL) Initiative (viz., www.adlnet.org) is to develop learning content as component units, called learning objects, that can be tagged, coded in a repository, identified, and reused or repurposed for other instructional applications. The SCORM, a reference model based on emerging e-learning industry standard specifications, permits interoperability of learning content and learning management systems. The model includes a specification, developed by the IMS Global Learning Consortium, for question and test interoperability (QTI). The specification defines a standard format that allows interoperability for questions and tests between different computer systems. Computer software that supports QTI will allow export into and import from this format, so that if you computerize questions or tests on one system, then the material will also be usable on another system. This allows exchange of questions between learning management systems, content authors, content libraries, and other forms of question pools.

The specification does not limit product designs by prescribing certain user interfaces, pedagogical paradigms, or policies that constrain innovative use of question-generation learning strategies. The QTI specification does afford the opportunity to capture in a standard format the questions generated during the course of distributed learning. Once in this format, they may find multiple uses to enhance a learning experience as described elsewhere in this report.

Summary

This report provides a rationale for question generation as a viable learning multiplier in training environments. The rationale was derived from a thorough review of recent research on questioning from multiple perspectives: psychology, information systems design, cognitive science, and computational linguistics. Based on this review, nine practices were identified for immediate use in both classroom and distributed learning settings. Furthermore, an application area that builds from a base of frequently asked questions during training (or even those posed at the workplace) was identified as a key focus for continuing experimentation and development.

If employed properly, question generation strategies in DL can increase a soldier's depth of understanding about the workings of a complex system. This is beneficial in terms of not only higher immediate learning outcomes but also in improved retention, as degree of original learning is the best single predictor of knowledge/skill retention (Wisher, Sabol, and Ellis, 1999). The advantages of question generation in training are particularly important to the Army. This is especially so as the Army transitions to the complexities of a digital battleforce while transforming its training delivery to a soldier-centric DL mode. A similar argument was made separately for the potential of collaborative learning tools in soldier-centric settings (Bonk and Wisher, 2000).

The provisions for answering the questions students pose in DL settings, particularly asynchronous DL, are not clear at this time. What is known is that the deeper the question and the more thoughtful the response the better the learning. The potential is great. Rather than being an afterthought to future training systems, a question generation mechanism should be investigated as an essential feature.

References

- Adate, V.U. (1998). Students generating test items: A teaching and assessment strategy. *The Mathematics Teacher*, 91, 198-202.
- Allen, J. (1983). Recognizing intentions from natural language utterances. In M. Brady & R.C. Berwick (Eds), *Computational models of discourse* (pp. 107-166). Cambridge, MA: MIT Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Baker, L. (1985). How do we know when we don't understand? Standards for evaluating text comprehension. In D.L. Forrest-Pressley, G.E. Mackinnon, T.G. Waller (Eds) *Metacognition, cognition and human performance* (pp. 155-205). New York: Academic Press.
- Beck, I.L., McKeown, M.G., Hamilton, R.L., & Kucan, L. (1997). *Questioning the Author: An approach for enhancing student engagement with text*. Delaware: International Reading Association.
- Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain*. New York: McKay.
- Bonk, C.J. & Wisher, R.A. (2000). Applying collaborative and e-learning tools to military distance learning: A research framework. Technical Report 1107. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Bransford, J. D., Goldman, S. R., & Vye, N. J. (1991). Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg & L. Okagaki (Eds.), *Influences on children* (pp. 147-180). Hillsdale, NJ: Erlbaum.
- Brown, A. L. (1988). Motivation to learn and understand: On taking charge of one's own learning. *Cognition and Instruction*, 5, 311-321.
- Burbules, N.C., & Linn, M.C. (1988). Response to contradiction: Scientific reasoning during adolescence. *Journal of Educational Psychology*, 80, 67-75.
- Cassell, J., & Thorisson, K.R. (1999). The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13, 519-538.
- Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.
- Ciardiello, A.V. (1998). Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy*, 42, 210-219.
- Clark (1996). *Using language*. Cambridge: Cambridge University Press.

Cohen, M.M., & Massaro, D.W. (1994). Development and experimentation with synthetic visible speech. *Behavior Research Methods, Instruments, and Computers*, 26, 260-265.

Collins, A. (1988). Different goals of inquiry teaching. *Questioning Exchange*, 2, 39-45.

Craig, S.D., Gholson, B., Ventura, M., Graesser, A.C., & the TRG (2000). Overhearing dialogues and monologues in virtual tutoring sessions: Effects on questioning and vicarious learning. *International Journal of Artificial Intelligence in Education*, 11, 242-253.

Francis, W.N., & Kucera, N. (1982). Frequency analysis of English usage. Houghton-Mifflin.

Department of Defense (1983). *Armed Services Vocational Aptitude Battery, Form 12a*. Washington, D.C.: Department of Defense.

Dillon, T.J. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.

Dillon, J. T., Golding, J., & Graesser, A. C. (1988). An annotated bibliography of question asking. *Questioning Exchange*, 2, 81-85.

Downing, S.M., & Haladyna, T.M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education*, 10, 61-82.

Edelson, D.C., Gordin, D.N., & Pea, R.D. (1999). Addressing the challenges of inquiry-based learning through technology and curriculum design. *The Journal of the Learning Sciences*, 8, 391-450.

Festinger, L. (1957). *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson.

Fishbein, H.D., Eckart, T., Lauver, E., Van Leeuwen, R., & Langmeyer, D. (1990). Learners' questions and comprehension in a tutoring setting. *Journal of Educational Psychology*, 82, 163-170.

Flammer, A. (1981). Towards a theory of question asking. *Psychological Research*, 43, 407-420.

Foltz, P.W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments*,

Foltz, P.W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, 8, 111-128.

Fox, B. (1993). The human tutorial dialog project. Hillsdale, NJ: Erlbaum.

Glenberg A.M., & Epstein, W. (1985). Calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11, 702-718.

Glenberg, A.M., Wilkinson, A.C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition*, 10, 597-602.

Good, T.L., Slavings, R.L., Harel, K.H., & Emerson, M. (1987). Students' passivity: A study of question asking in K-12 classrooms. *Sociology of Education*, 60, 181-199.

Graesser, A. C., Gordon, S. E., & Brainerd, L. E. (1992). QUEST: A model of question answering. *Computers and Mathematics with Applications*, *23*, 733-745.

Graesser, A. C., Lang, K., & Horgan, D. (1988). A taxonomy of question generation. *Questioning Exchange*, *2*, 3-16.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 411-436). Orlando, FL: Academic Press.

Graesser, A. C., & McMahan, C. L. (1993). Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology*, *85*, 136-151.

Graesser, A.C., Olde, B., & Lu, S. (2001). Question-driven explanatory reasoning about devices that malfunction. In T. Filjak (Ed.), *Proceedings of the 36th International Applied Military Psychology Symposium*.

Graesser, A.C., Olde, B., Pomeroy, V., Whitten, S., Lu, S., & Craig, S. (in press). Inferences and questions in science text comprehension. In book edited by J. Otero and M. Helena (Eds.), *Science text comprehension*.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104-137.

Graesser, A.C., Person, N., Harter, D., & TRG (2000). Teaching tactics in AutoTutor. *Proceedings of the workshop on modeling human teaching tactics and strategies at the Intelligent Tutoring Systems 2000 conference*. University of Quebec at Montreal, 49-57.

Graesser, A. C., Person, N., & Huber, J. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems*. Hillsdale, NJ: Erlbaum.

Graesser, A.C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (in press). Intelligent tutoring systems with conversational dialogue. *AI Magazine*.

Graesser, A.C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the TRG (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, *1*, 35-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, *8*, 129-148.

Greer, J. (1999). Computer support for peer-help networks. *Proceedings of the International Conference on Computers in Education* (pp. 39-44). IOS Press.

Hacker, D.J., Dunlosky, J., & Graesser, A.C. (1998)(Eds.). *Metacognition in educational theory and practice*. Mahwah, NJ: Erlbaum.

Hillman, D.C.A. (1999). A new method for analyzing patterns of interaction. *The American Journal of Distance Education*, 13, 37-47.

Johnson, D.W., & Johnson, R.T. (1989). *Cooperation and competition: Theory and research*. Edina, MN: Interaction Book Company.

Johnson, W. L., & Rickel, J. W., & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47-78.

Kagan, S. (1994). *Cooperative learning*. San Juan Capistrano, CA: Resources for Teachers.

Kerry, T. (1987). Classroom questions in England. *Questioning Exchange*, 1, 32-33

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14, 366-381.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Educational Research Journal*, 29, 303-323.

King A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31, 338-368.

King, A., & Rosenshine, B. (1993). Effects of guided cooperative-questioning on children's knowledge construction. *Journal of Experimental Education*, 6, 127-148.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kreuz, R. J., & Graesser, A. C. (1993). The assumptions behind questions in letters to advice columnists. *Text*, 13, 65-89.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Lauer, T., Peacock, E., & Graesser, A. C. (1992) (Eds.). *Questions and information systems*. Hillsdale, NJ: Erlbaum

Lehmann, F. (1992)(Eds.) *Semantic networks in artificial intelligence*. New York: Pergamon.

Lehnert, W.G. (1997). Information extraction: What have we learned? *Discourse Processes*, 23, 441-470.

MacGregor, S.K. (1988). Use of self-questioning with a computer-mediated text system and measures of reading performance. *Journal of Reading Behavior*, 20, 131-148.

Maki, R.H. (1998). Text predictions over text material. In D.J. Hacker, J Dunlosky, & A.C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp 117-145). Mahwah, NJ: Erlbaum.

Martinez, M.E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207-218.

Messeck, S. (1987). *Assessment in schools: Purposes and consequences* (RR-87-51). Princeton, NJ: Educational Testing Service.

National Council of Teachers of English (NCTE) (1996). *Standards for the English language arts*. Urbana, IL: National Council of Teachers of English.

National Council of Teachers of Mathematics (NCTM) (1989). *Curriculum and Evaluation Standards for school mathematics*. Reston, VA.

National Research Council (NRC)(1996). National science education standards. Washington, D.C.: National Academy Press.

Ogata, H., Sueda, T., Furugori, & Yano, Y. (1999). Augmenting collaboration beyond classrooms through online social networks. *Proceedings of the International Conference on Computers in Education* (pp. 39-44). IOS Press.

Otero, J., & Campanario, J.M. (1990). Comprehension evaluation and regulation in learning from science texts. *Journal of Research in Science Teaching*, 27, 447-460

Otero, J., & Graesser, A.C. (in press). PREG: Elements of a model of question asking. *Cognition & Instruction*.

Palinscar, A. S., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117-175.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York: Basic Books.

Pressley, M., & Forrest-Pressley, D. (1985). Questions and children's cognitive processing. In A.C. Graesser, & J.B. Black (Eds.), *The psychology of questions* (pp. 277-296). Hillsdale, NJ: Erlbaum.

Rickel, J., & Johnson, W.L. (1999). Animated agents for procedural training in virtual reality: Perception, cognition, and motor control. *Applied Artificial Intelligence*, 13, 343-382.

Robertson, S.P. (1994). TSUNAMI: Simultaneous understanding, answering, and memory interaction for questions. *Cognitive Science*, 18, 51-86.

Rogoff, B. (1990). *Apprenticeship in Thinking*. New York: Oxford University Press.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, 66, 181-221.

- Scardamalia, M., & Bereiter, C. (1985). Fostering the development of self-regulation in children's knowledge processing. In S.F. Chipman, J.W. Segal, & R. Glaser (Eds.), *Thinking and learning skills*, Vol. 2 (pp. 563-577). Hillsdale, NJ: Erlbaum.
- Schank, R.C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.
- Schank, R.C. (1999). *Dynamic memory revisited*. Cambridge: Cambridge University Press.
- Sebrechts, M.M., & Swartz, M.L. (1991). Question asking as a tool for novice computer skill acquisition. *Proceedings of the 1991 International Conference on Human Computer Interaction* (pp. 293-297).
- Slavin, R.E. (1995). *Cooperative learning: Theory, research, and practice* (Ed. 2). Boston: Allyn & Bacon.
- Songer, N.B. (1996). Exploring learning opportunities in coordinated network-enhanced classrooms: A case of kids as global scientists. *The Journal of the Learning Sciences*, 5, 297-327.
- Springer, L., Stanne, M.E., & Donovan, S.S. (1999). Effects of small-group learning on undergraduates in science, mathematics, engineering, and technology: A meta-analysis. *Review of Educational Research*, 69, 21-51.
- Van der Meij, H. (1987). Assumptions of information-seeking questions. *Questioning Exchange*, 1, 111-118.
- VanLehn, K., Jones, R.M., & Chi, M.T. (1992). A model of the self-explanation effect. *The Journal of the Learning Sciences*, 2, 1-59.
- Vrasidas, C., & McIsaac, M.S. (1999). Factors influencing interaction in an on-line course. *The American Journal of Distance Education*, 13, 22-36.
- Vygotsky, L.S. (1978). *Mind in society*. Cambridge, MA: Harvard University Press.
- Webb, N.M., Troper, J.D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Experimental Psychology*, 87, 406-423.
- Webber, B. (1988). Question answering. In S.C. Shapiro (Ed.), *Encyclopedia of artificial intelligence*: Vol. 2 (pp. 814-822). New York: Wiley.
- Wiemer-Hastings, P., Wiemer-Hastings, K., and Graesser, A. (1999). Improving an intelligent tutor's comprehension of students with Latent Semantic Analysis. In S.P. Lajoie and M. Vivet, *Artificial Intelligence in Education* (pp. 535-542). Amsterdam: IOS Press.
- Wisher, R.A., Champagne, M.V., Pawluk, J.L., Eaton, A., Thornton, & Curnow, C.K. (1999). Training and distance learning: An assessment of research findings. Technical Report #1095, US Army Research Institute for the Behavioral and Social Sciences. Alexandria, VA.

Wisher, R.A., & Curnow, C.K. (1999). Perceptions and effects of image transmissions during internet-based training. *The American Journal of Distance Education*, 13, 37-51.

Wisher, R.A., Sabol, M.A., & Ellis, J. (1999). Staying sharp: The Retention of Military Knowledge and Skills. Special Report #39, US Army Research Institute for the Behavioral and Social Sciences. Alexandria, VA

APPENDIX A

Question Taxonomy Proposed by Graesser and Person (1994).

The table below presents the 18 questions categories that are based on question content and that were sincere information-seeking (SIS) questions.

QUESTION CATEGORY EXAMPLES	GENERIC QUESTION FRAMES AND
1. Verification	Is X true or false? Did an event occur? Does a state exist?
2. Disjunctive	Is X, Y, or Z the case?
3. Concept completion	Who? What? When? Where?
4. Feature specification	What qualitative properties does entity X have?
5. Quantification	What is the value of a quantitative variable? How much? How many?
6. Definition questions	What does X mean?
7. Example questions	What is an example or instance of a category?).
8. Comparison	How is X similar to Y? How is X different from Y?
9. Interpretation	What concept or claim can be inferred from a static or active pattern of data?
10. Causal antecedent	What state or event causally led to an event or state? Why did an event occur? Why does a state exist? How did an event occur? How did a state come to exist?
11. Causal consequence	What are the consequences of an event or state? What if X occurred? What if X did not occur?
12. Goal orientation action?	What are the motives or goals behind an agent's
13. Instrumental/procedural	Why did an agent do some action? What plan or instrument allows an agent to accomplish a goal? How did agent do some action?
14. Enablement	What object or resource allows an agent to accomplish a goal?
15. Expectation	Why did some expected event <u>not</u> occur? Why does some expected state <u>not</u> exist?
16. Judgmental	What value does the answerer place on an idea or advice?
17. Assertion	What do you think of X? How would you rate X? A declarative statement that indicates the speaker does not understand an idea.
18. Request/Directive	The questioner wants the listener to perform some action.

APPENDIX B

TRADOC Guidelines for Preparing Multiple Choice Test Items

Lessons Learned for All Test Items

- Write the items in preliminary form during the instructional system development period.
- Use a test blueprint or outline to keep an appropriate relationship between the items on the test and the instructional objectives.
- Base each test item on an important point, idea, or skill.
- Write items to measure understanding or ability to apply principles.
- Test one, and only one, point or idea per test item.
- Write items that require specific knowledge of material studied, not items that require general knowledge or experience.
- Use clear and concise language that is appropriate for the conceptual difficulty level of the specific objective being tested.
- Present each test item task as simply and straightforwardly as possible.
- Keep test items free of extraneous, ambiguous, or confusing material.
- Keep test items free of tricky expressions, slang, or other tricky requirements.
- Review test items from other sources, such as textbooks, and other instructors.
- Use original language, not that found in textbooks or other instructional materials for the course.
- Eliminate any clues within the test item, or clues that relate to other items in the test.
- Be especially sensitive to clues or suggestions that could help a naive examinee (one who does not have the knowledge or skill that should be able to answer the item correctly).
- For tests that measure discrimination, concrete concept, or defined concept intellectual skills, make each test item independent of other items. Ensure that the answer to one test item is not dependent on the answer to other test items. (Some tests that measure rule learning or problem solving intellectual skills, verbal information skills, cognitive strategy skills, or the memorization component of psychomotor skills may be designed to require correct answers to a sequence of test items. These tests require that the correct answer to one, or a series of test items, is dependent on the correct answers to other previous test items).
- Ensure that test items are reviewed by other instructors and content specialists to help eliminate ambiguity, technical errors, or other errors in the test item.
- Ensure test items are reviewed by individuals who are not content specialists for ambiguity, clues for naive examinees, and for selected-response test items, plausibility for the naive examinee.
- Ensure that the test item has "face validity", measures a specific objective, and relates to the content studied in the course of instruction.
- Avoid the appearance of bias in the test item (e.g., race, gender, cultural, ethnic, regional, handicapped, age-group, or other apparent bias).
- Construct test items that have a clearly correct or clearly best answer.
- Follow standard rules of punctuation and grammar.
- For test items based on an opinion or authority, state whose opinion or what authority.
- Do not require unnecessarily exact or difficult operations. Test items should match objective criterion standards.
- Do not use specific determiners such as "always", "never", "none", and "all" in test items.
- Restrict the number of different item formats in a test. Use the most valid formats. Group items in the same format together.
- Use scenarios, pictorial material, or other graphics only when they are relevant to an objective or topic measured by the test item, and only when required for the test item to effectively measure

an intellectual or psychomotor skill.

- When a scenario, picture, or graphic is used, provide specific test item directions referring to it.
- If scenarios are used for a test item, ensure that they are realistic and appropriate for the test item.
- If pictorial material or other graphics are used for a test item, ensure that they are clearly drawn and labeled.

Lessons Learned for Multiple-Choice Test Items

- Write the stem so it clearly defines the test item task. Word the stem so that the examinee knows what is required without seeing the response options. Generally, writing the stem as a question helps to set the test item task more clearly.
- When the stem is written as an incomplete statement, the option statements should complete the sentence, rather than beginning the item stem, or being inserted in the middle of the item stem.
- Reduce the “reading load” as much as possible. Avoid repeating words in the option statements, by placing these words in the stem.
- Do not provide verbal clues that point to the correct option or to elimination of incorrect option(s), such as disagreement between singular or plural, “a” and “an”, etc.
- Make all option statements fit or match the stem.
- Have one, and only one, correct option.
- Make the options approximately equal in length. Avoid the tendency to make the correct option more detailed.
- Make the options logically parallel, and about equal in complexity.
- Make the options grammatically and syntactically parallel. Use grammatically and syntactically parallel words in the stem, the distracters, and the correct option.
- Avoid using modifiers such as “sometimes” and “usually” in the options.
- Ensure that each option has a unique meaning. Eliminate distracters with the same or similar meanings from the test item.
- Make all distracters plausible to a naive examinee. Do not include implausible or impossible options as distracters in a test item.
- Arrange the options in some appropriate, logical order.
- Vary the position of the correct option.
- Avoid using “all of the above” as an option, and use “none of the above” sparingly.
- Avoid using negative words, including “except” in the stem and in the options. If it is necessary to use negative words, underline, capitalize, or highlight them for emphasis and examinee visibility.